ФІЛЬТРУВАННЯ ІНТЕРНЕТ СПАМУ ЗА ДОПОМОГОЮ ОБРОБКИ ПРИРОДНОЇ МОВИ

Філоненко О.В., Черних О.П., Шеін О.М.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

Електронна інформація у наш час − дуже великий ринок, на якому багато як якісного і потрібного контенту, так не якісного і шкідливого. У наші поштові скриньки надходить багато листів, деякі офіційні, деякі із рекламою, деякі із масових розсилок (спаму), а також із іншими видами нотифікацій. Таким чином, питання класифікації листів, а також фільтрування спам-розсилок є дуже актуальним.

Існує багато способів фільтрування спам-розсилок, деякі більш ефективні, деякі менш. В даній роботі увага приділяється способу фільтрування за допомогою обробки природної мови. Він є досить ефективним і добре адаптується до нових видів спаму.

Для вирішення наведеної проблеми було обрано платформу StanfordNLP та мову програмування C#. Платформа StanfordNLP дозволяє використовувати методи машинної обробки текстів для класифікації та відсіювання шкідливих листів. Бібліотека містить багато модулів, але корисними для нашої задачі будуть класифікатор (Classifier) та інструмент для встановлення зв’язків між поняттями (Relation Extractor). Класифікатор потрібен для швидкої класифікації листів за типами (наприклад, реклама, загальні, запрошення, нотифікації та інші). Він буде грати роль розподілювача загального потоку і відокремлювати з нього можливі рекламні та шкідливі листи для подальшого аналізу. Наступним буде застосування «витягувала» залежностей, який порівнюватиме залежності між частинами речень у листі із вже відомими прикладами залежностей у спам-розсилках та шкідливих листах. Таким чином, будуть ефективно фільтруватись і ті та інші. Система вимагатиме тренування, яке може бути здійснене на деякому порівняно невеликому сеті даних. Сет поділяється на дві частини – тренувальну (основну) та перевірочну (невеликого розміру). Потім моделі тренуються тренувальним сетом і перевіряються на перевірочному сеті. Після успішного тренування буде отримано моделі, що дозволяють фільтрувати вхідні листи за їх приналежністю до вже відомої класифікації. А завдяки гнучким моделям нові загрози і спам-листи будуть із великою вірогідністю також відфільтровані.

Для розробника платформа StanfordNLP цікава тим, що надає зручну базу для оперування із природної мовою, а на цій базі можна зробити будь-яку скільки завгодно складну систему. Наприклад, можливо визначити рекламні листи та автоматично помічати їх як рекламу, шкідливі − видаляти одразу, а спам-розсилання переміщувати одразу до теки «спам».